# Semi-supervised Learning Models

# Outline

- Semi-supervised learning with consistency regularization
  - Supervised + consistency regularization (unsupervised)
  - 2 forms Consistency regularization: data augmentation and teacher-student
- $\Gamma$-Model, $\Pi$-Model (Ramsmus et al., NIPS 2015; Sajjadi et al., NIPS 2016), simple data augmentation
- Universal data augmentation (UDA) (Xie et al., 2019), complex data-augmentation
- Temporal Ensemble (Laine et al., ICLR 2017), teacher-student
- Mean Teachers (Tarvainen et al., NIPS 2017), teacher-student
- FixMatch (Sohn et al., 2020), simple-complex data-augmentation + teacher-student
- Towards NLP

# Semi-supervised Learning with Consistency Regularization

- Supervised part

$$L(x_l, y_l, \theta) = \mathrm{CE}(q(y_l|x_l), p(y|x_l;\theta)$$

- Consistency regularization part (a classifier should give consistency output for similar data points (invariant semantics, changing as many pixels of an image without changing its meaning)): 2 forms
  - data augmentation: the two inputs are similar to each other $x_{\mathrm{ul}} \sim x_{\mathrm{ul}}^+$,

$$L_c(x_{\mathrm{ul}}, \theta) = \mathcal{J}(p(y|x_{\mathrm{ul}};\theta),\ p(y|x_{\mathrm{ul}}^+;\theta))$$

  - teacher-student: the student $p$ tries to match the teacher $p^+$'s prediction, teacher and student have similar classifiers $p \sim p^+$, e.g., same architecture but different parameters

$$L_c(x_{\mathrm{ul}}, \theta) = \mathcal{J}(p(y|x_{\mathrm{ul}};\theta),\ p^+(y|x_{\mathrm{ul}};\theta^+))$$

- Note there is no clear distinction between the two forms:
  - both $x_{\mathrm{ul}}$ and $x_{\mathrm{ul}}^+$ can be augmentation of the same input
  - classifier $p$ can also apply data augmentation
  - if the classifier $p^+$ results in a fixed discrete pseudo-label or continuous distribution (and is not back-propagated) then the method belongs to teacher-student form, else the method is considered data augmentation

# $\Gamma$-Model and $\Pi$-Model, and UDA – Data Augmentation

$$L_c(x_{\mathrm{ul}}, \theta) = \mathcal{J}(p(y|x_{\mathrm{ul}};\theta),\ p(y|x_{\mathrm{ul}}^{+};\theta))$$

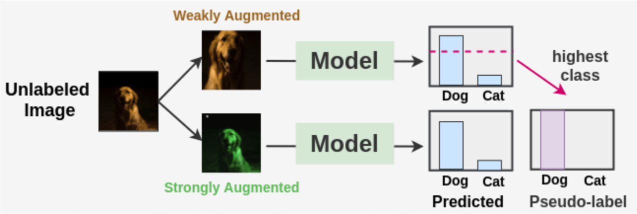
- $x_{\mathrm{ul}}$ and $x_{\mathrm{ul}}^{+}$ are both augmentation of the same inputs
  - explicit augmentation: for images, invariant transformations such as random crop, flip, rotate, cutout images, changing brightness, color, contrast
  - implicit augmentation: model's internal stochasticity such as dropout (different passes have different dropouts thus produce different outputs), virtual adversarial examples, mix-up (strange but interesting idea)
- $\Gamma$-Model and $\Pi$-Model apply simple augmentation: dropout + random crop and flip images
- (Universal data augmentation) UDA applies a complex reinforcement learning strategy to find the best set of augmentations out of 16 augmentation choices (+ their parameters) for each image domain.

# Temporal Ensemble and Mean Teacher – teacher-student

$$L_c(x_{\mathrm{ul}}, \theta) = \mathcal{J}(p(y|x_{\mathrm{ul}}; \theta),\ p^+(y|x_{\mathrm{ul}}; \theta^+))$$

- Temporal Ensemble
  - output $Z$ of $p^+$ of an input is an accumulated prediction of $p$ of that input
  - $p^+ = Z = \alpha Z + (1 - \alpha)z$, where $z$ is the current prediction,
  - $Z$ is first initialized to be 0
- Mean Teacher
  - parameters $\theta^+$ of $p^+$ is an accumulated parameters of $p$
  - $\theta^+ = \alpha\theta^+ + (1 - \alpha)\theta$, where $\theta$ is the current parameters
  - $\theta^+$ is first initialized to be 0

# FixMatch – simple-complex data-augmentation + teacher-student




- teacher classifier
  - apply simple data augmentation
  - the classifier's output is the class with highest probability (larger than some threshold)
- student classifier:
  - apply complex data augmentation, i.e., RL strategy
  - student tries to match output of the teacher classifier.

# Towards NLP

- Teacher-student seems "easy" to apply
  - Single-/Multi-source cross-lingual NER via Teacher-Student Learning (Wu et al., ACL 2020)
- Data augmentation
  - implicit augmentation: dropout, virtual adversarial examples, ... they work but not as good as explicit augmentation
  - explicit augmentation:
    - it is not easy to do since words are discrete, if we change few words we may change the semantics
    - random noise injection: embeddings noise, spelling error, unigram noising, ...
    - lexical substitutions (wordnet, word-embeddings, masked language model),
    - back translation: translate sentences into different languages then translate them back to original language
    - generative methods

# Current Benchmarks on Text Classifications

| Fully supervised baseline | | | | | | | |
|---|---|---|---|---|---|---|---|
| **Datasets** (# Sup examples) | | IMDb (25k) | Yelp-2 (560k) | Yelp-5 (650k) | Amazon-2 (3.6m) | Amazon-5 (3m) | DBpedia (560k) |
| Pre-BERT SOTA | | *4.32* | 2.16 | 29.98 | 3.32 | 34.81 | 0.70 |
| $\text{BERT}_{\text{LARGE}}$ | | 4.51 | *1.89* | *29.32* | *2.63* | *34.17* | *0.64* |

| Semi-supervised setting | | | | | | | |
|---|---|---|---|---|---|---|---|
| **Initialization** | **UDA** | IMDb (20) | Yelp-2 (20) | Yelp-5 (2.5k) | Amazon-2 (20) | Amazon-5 (2.5k) | DBpedia (140) |
| Random | ✗ | 43.27 | 40.25 | 50.80 | 45.39 | 55.70 | 41.14 |
| | ✓ | 25.23 | 8.33 | 41.35 | 16.16 | 44.19 | 7.24 |
| $\text{BERT}_{\text{BASE}}$ | ✗ | 18.40 | 13.60 | 41.00 | 26.75 | 44.09 | 2.58 |
| | ✓ | 5.45 | 2.61 | 33.80 | 3.96 | 38.40 | 1.33 |
| $\text{BERT}_{\text{LARGE}}$ | ✗ | 11.72 | 10.55 | 38.90 | 15.54 | 42.30 | 1.68 |
| | ✓ | 4.78 | 2.50 | 33.54 | 3.93 | 37.80 | 1.09 |
| $\text{BERT}_{\text{FINETUNE}}$ | ✗ | 6.50 | 2.94 | 32.39 | 12.17 | 37.32 | - |
| | ✓ | **4.20** | **2.05** | **32.08** | **3.50** | **37.12** | - |

*Thank you !*

# Current Benchmarks on Image Classifications

| | CIFAR-10 | | |
|---|---|---|---|
| Method | 40 labels | 250 labels | 4000 labels |
| Π-Model | - | 54.26±3.97 | 14.01±0.38 |
| Pseudo-Labeling | - | 49.78±0.43 | 16.09±0.28 |
| Mean Teacher | - | 32.32±2.30 | 9.19±0.19 |
| MixMatch | 47.54±11.50 | 11.05±0.86 | 6.42±0.10 |
| UDA | 29.05±5.93 | 8.82±1.08 | 4.88±0.18 |
| ReMixMatch | **19.10**±9.64 | **5.44**±0.05 | 4.72±0.13 |
| FixMatch (RA) | **13.81**±3.37 | **5.07**±0.65 | **4.26**±0.05 |
| FixMatch (CTA) | **11.39**±3.35 | **5.07**±0.33 | **4.31**±0.15 |